Image Classification Based on CodeBook on CodeBooks

Katsuyuki TANAKA[1] Tetsuya TAKIGUCHI[2] Yasuo ARIKI[2]

1. Introduction

The Bag of Features (BoF) method is one of the most popular methods to provide the state-of-the-art in image classification performance. One of the important issues that affect the performance of image classification is the quality of the codebook and coding scheme. The simplest codebook generation and coding is a hard vector quantization (HVQ) method[1]. In HVQ, codebook is leaned by K-means (each centroid is a visual word) and each descriptor is encoded to the closest visual word (only one non-zero element per code). Wang[3] (LLC) successfully improved the image classification performance by incorporating locality constraint in coding.

Even though, some local descriptors are more informative to represent the characteristics of a particular concept of an image/object, the unsupervised way of generating codebook does not take such class information into account with the codebook. Futhermore, There are no clear standards of how many local descriptors per class/image should be sampled to generate a codebook.

In this paper, we propose a novel image classification approach, Locality-constrained Linear Coding with codebook on codebooks. The flow of our proposed method is, i) generate a class codebook from each class using local descriptors of the class, ii) generate a global codebook based on class codebooks, and iii) encode local descriptors to codes with LLC based on the global codebook. Since the algorithm does not go beyond the Euclidean distance based BoF approach, it is simple and easy to re-implement. Moreover, there no longer requires to try different sampling strategy to tune a codebook.

2. Related Works and Notations

Let $X = \{x_m \in \mathbb{R}^D\}_{m=1..M}$ be a set of D-dimensional local descriptors x_m extracted at M locations from an image I. The main concern of a codebook generation and coding is how to generate a codebook with K visual words $V = \{v_k \in \mathbb{R}^D\}_{k=1..K}$ and define the scheme to encode x_m to a K-dimensional code u_m (effectively $U = \{u_m \in \mathbb{R}^K\}_{m=1..M}$ is obtained by converting X).

[1] Faclity of Economics, Kobe University, 2-1, Rokkodai, Nada, Kobe, 657-8501, Japan
[2] Organization of Advanced Science and Technology, Kobe University, 1-1, Rokkodai, Nada, Kobe, 657-8501, Japan

In HVQ[1], visual words for the codebook are usually centroids learned by K-means. By assigning the nearest visual word v_k to x_m, a local descriptor x_m is encoded to a code u_m which only has one non-zero element ($Card(u_m) = 1$).

LLC[3] improved the quantization loss of a local descriptor by assigning a small combination of visual words to a code. In LLC, the choice of visual words to encode x_m is based on locality constraint. LLC assigns similar visual words v_k to x_m. Effectively, it enables to retain the correlations between similar local descriptors in codes.

$$\min_{U,V} \sum_{m=1}^{M} \|x_m - V u_m\|_2^2 + \lambda \|d_m \odot u_m\|_2^2 \qquad (1)$$

$$s.t. 1^T u_m = 1, \forall m$$

$$d_m = \exp\left(\frac{[dist(x_m, v_1), \cdots, dist(x_m, v_K)]^T}{\sigma}\right) \qquad (2)$$

The second term in Eq.(1) is the locality constraint, which enforces $d_m \in \mathbb{R}^K$ to u_m, where $\odot$ denotes the element-wise multiplication. Eq.(2) is the calculation of d_m. It calculates Euclidean distance between x_m and v_k ($dist(x_m, v_k)$) and it penalises visual words far from x_m, where σ is a parameter controls the speed of weight decay of d_m. In [3], Eq.(2) is solved by using kNN visual words of x_m to speed up the algorithm and reduce computation cost.

3. Proposed Method

Suppose C_n is the one of N image classes. Firstly, apply K-means on local descriptors $X_n = \{x_m \in \mathbb{R}^D, C_n\}_{m=1..M_{c_n}}$ which extracted only from the images with the same class C_n and obtain cK centroids $CV_n = \{cv_{n,ck} \in \mathbb{R}^D\}_{ck=1..cK}$ as described in Eq.(3). It effectively means to generate a class codebook for each class C_n with cK visual words. We call $cv_{n,ck}$ as a class visual word of the class C_n.

$$CV_n = \min_{CV} \sum_{m=1}^{M_{c_n}} \min_{ck=1...cK} \|x_m - cv_{n,ck}\|_2^2 \qquad (3)$$

Secondly, generate another codebook GV (global codebook) with gK visual words $GV = \{gv_{gk} \in \mathbb{R}^D\}_{gk=1..gK}$ by applying K-means on the $cK \times N$ class visual words, all class codebooks obtained by the previous step.

$$GV = \min_{GV} \sum_{n=1}^{N} \sum_{ck=1}^{cK} \min_{gk=1\cdots gK} \|cv_{n,ck} - gv_{gk}\|_2^2 \qquad (4)$$

Finally, LLC coding with the global codebook GV can be formulated as Eq.(5). We call this LLC with (global) codebook on (class) codebooks, ccLLC in short.

$$\min_{U} \sum_{m=1}^{M} \|x_m - GVu_m\|_2^2 + \lambda \|d_m \odot u_m\|_2^2 \qquad (5)$$

$$s.t. 1^T u_m = 1, \forall m$$

$$d_m = \exp\left(\frac{[dist(x_m, gv_1), \cdots, dist(x_m, gv_{gK})]^T}{\sigma}\right) \qquad (6)$$

A generating class codebook can be considered as extracting the representative local descriptors of the class, and find the relationship among these local descriptors by generating global codebook. Effectively, it tries to retain class information by generating a global codebook on top of class codebooks.

The computation cost of K-means iteration becomes critical with big dataset, since the number of both on inter-class and intra-class local descriptors dramatically increase. Though the common practice of reducing the cost of solving a big linear system is randomly sample a sbuset of all local descriptors available, it becomes difficult which local descriptors should be sampled to absorb the charactristics of classes with big dataset. The proposed approach only requires one parameter cK to control sampling rate and it is fixed to all classes. Despite the cost of K-means per class, it is also possible to break the cost of solving big linear system to solving the number of smaller linear systems and it is easy to implement.

4. Experiments

4.1 Experiment Conditions

We conduct the experiments on Caltech-101 and 15-Scenes to evaluate the performance of our proposed method. All images are resized to fit the 300×300 pixels box by keeping the aspect ratio. We use SIFT as a local descriptor and they are extracted densely over the grayscale images with four scales [1, 2, 4, 8] from every 8 pixels. We train the class codebooks with $cK = 2048$ and the global codebook with $gK = 1024$, encoding is performed by LLC with the global codebook. The codes are pooled with three levels of pyramids (1, 2, 4)[2] and max pooling is used for all experiments. One-vs-rest classification is performed on linear SVM for each class. All experiments are run 10 times over 10 random splits of training and test data. We report the mean accuracy and standard deviation of these runs.

The performance of the proposed method is compared with LLC. K-means is performed over total of 400,000 (400k) and 800,000 (800k) evenly sampled local descriptors per class. To provide comprehensive analysis, the experiments are also conducted on codebook on codebooks with HVQ (ccHVQ) and compare with normal HVQ.

Table 1 Average classification rate on Caltech-101 and 15-Scenes

Algorithm	Caltech-101(%)	15-Scenes(%)
HVQ(400k)	73.45±0.83	80.02±0.61
HVQ(800k)	73.31±0.89	79.77±0.68
LLC(400k)	**75.41±1.08**	81.29±0.60
LLC(800k)	75.09±0.98	81.19±0.59
ccHVQ	73.42±1.13	80.05±0.74
ccLLC	75.20±1.33	**81.35±0.56**

4.2 Experiment

The experiments on Caltech-101 are conducted by following the standard experimental setup, we use randomly selected 30 images for training and the remaining for testing (no more than 50 images). Middle column of Table.1 shows the results.

The performance of the proposed method is comparable to baselines, it measured 73.42% on ccHVQ and 75.2% on ccLLC. The classification rate is between 400k and 800k sampled baselines. Surprisingly, 800k sampled showed the lower classification rate than 400k sampled. As far as the number of local descriptors used to generate the codebook, the proposed method only required 2048 local descriptors per 102 classes to provide comparable performance to baselines.

We conducted further experiment on 15-Scenes. By following the standard setup, we use 100 images for training and the rest for testing and results are shown in right column of Table1.

The classification rate of the proposed method is 80.05% on ccHVQ and 81.35% on ccLLC. They are comparable performance to baselines, and again increasing sampling rate from 400k to 800k does not improve the performance.

5. Conclusion

In this paper, we proposed to generate a global codebook based on class codebooks generated from each class and encode with Locality-constrained Linear Coding. From the evaluation, generating the global codebook using the class visual words showed the comparable classification performance with existing approaches. The proposed method is furthur evaluated by applying combination of various existing codebook generation on class codebook and global codebook with various different datasets.

References

[1] Csurka, G., Dance, C. R., Fan, L., Willamowski, J. and Bray, C.: Visual categorization with bags of keypoints, *Workshop on Statistical Learning in Computer Vision, ECCV'04*, pp. 1–22 (2004).

[2] Lazebnik, S., Schmid, C. and Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, *CVPR'06*, pp. 2169 – 2178 (2006).

[3] Wang, J., Yang, J., Yu, K., Lv, F., Huang, T. and Gong, Y.: Locality-constrained Linear Coding for image classification, *In CVPR'10*, pp. 3360–3367 (2010).